

Data and Information Quality at CIHI: a Foundation for Meaningful Health Information

Presenters:

Heather Richards and Chad Gyorfi-Dyke
(Data Quality, CIHI)

CIHI: Who We Are

- A national, independent, non-profit agency
- Mandate:
 - national coordination mechanism for health information in Canada
 - provide accurate and timely information for:
 - sound health policy
 - effective management of the health system
 - public awareness of health determinants

CIHI: What We Do

- Collect, process and maintain data for a growing number of national and provincial health databases and registries
- Coordinate/promote development and maintenance of health information standards
- Identify and develop priority health indicators at the national, provincial, and regional levels
- Produce analytical products including annual reports and special studies

Data and information quality at CIHI

- Intrinsic to mandate
 - Inform public policy
 - Support health care management
 - Build public awareness about the factors that affect health

- Engage in rigorous activities to ensure data collected and provided is the highest standard



DATA**QUALITY**

Corporate strategy

- Complete data quality program
- Continuously improve data quality within CIHI and broader health sector
- Initiatives aimed at **prevention**, **early detection** and **resolution** of data quality issues

CIHI's Data Quality Strategy

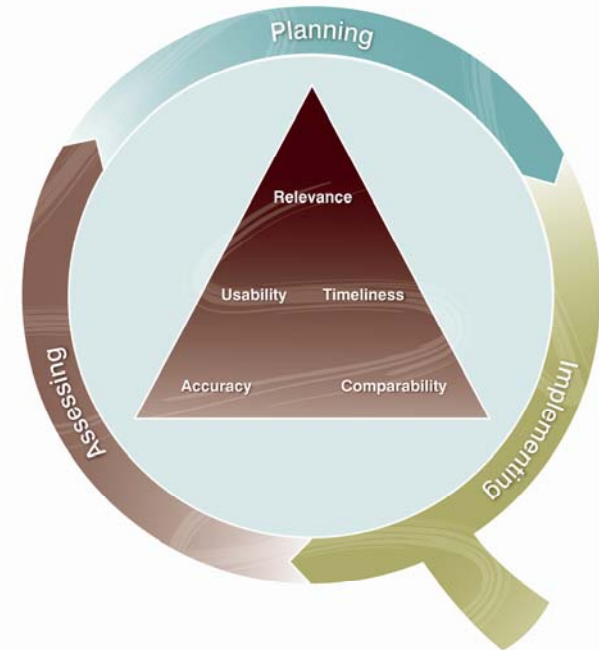


Defining data quality

CIHI's Data Quality Framework

There are five dimensions that comprise CIHI's strategy for an ongoing approach to maintaining data quality:

- **Accuracy**
- **Comparability**
- **Timeliness**
- **Usability**
- **Relevance**



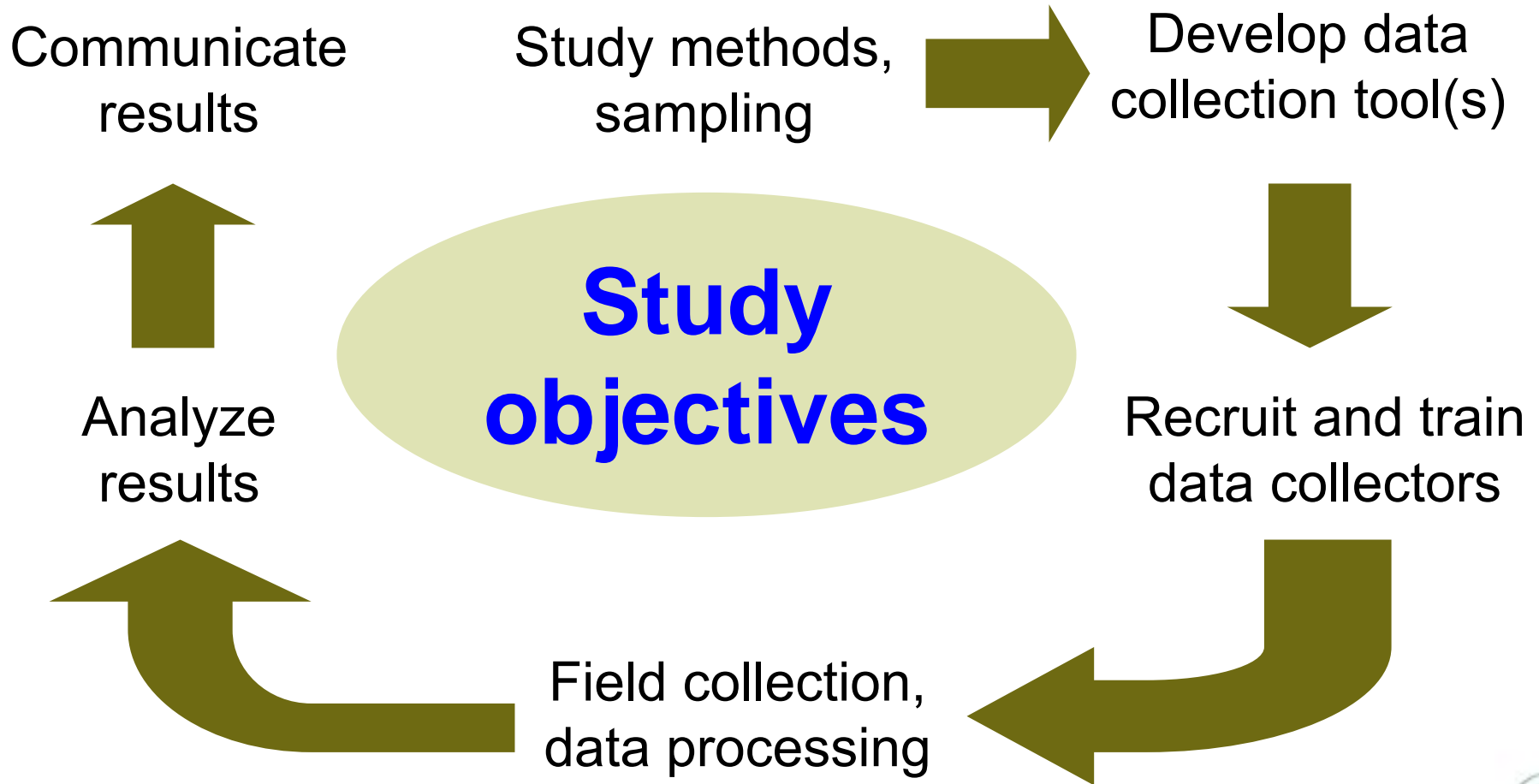
For more information, go to www.cihi.ca

Some examples of CIHI data quality activities

- Develop tools to identify strengths and gaps in the data
- Commission special data quality studies
- Conduct data mining analysis
- Build meta data repository
- Data quality reporting to provinces and territories
- Educational offerings
- Collaborate with data suppliers, vendors and other stakeholders
- Publish standards for clinical and financial reporting
- Participate on provincial/territorial advisory committees and data quality committees

Special Studies at CIHI

CIHI data quality studies



Formulating the statement of objectives

ref: Statistics Canada, Survey Methods and Practices

“Begin with the end in mind” Stephen R. Covey

- Determine...
 - Information needs
 - Users and uses of the data
 - Main concepts and operational definitions
 - The study content
 - The analysis plan (e.g. proposed tabulations)
- Can be constrained by... time, resources, politics, confidentiality, response burden

Example – Canadian Organ Replacement Register (CORR)

- National database on vital organ replacement therapy in Canada, with the goal of enhancing treatment, research and patient care.
- Developed in the 1970's and in 1995 was transferred to CIHI.
- Its mandate is to record and analyze the level of activity and outcomes of solid organ transplantation and renal dialysis activities.



Example – Canadian Organ Replacement Register (CORR)

- Concern with the quality of data for “new” dialysis patients
- Risk factor data key for analysis
- Gaps in the CORR instruction manual
- Suspect facility staff have different practices for completing forms

SECTION E—PRIMARY DIAGNOSIS AND RISK FACTOR HISTORY

| | | | |
|--|--------------------------|--------------------------|--------------------------|
| Primary renal disease (see codes on page 3) | | | |
| Specify _____ | | | |
| Risk Factors/Comorbid Conditions (Check one response per condition.) | | | |
| | No | Yes | Unknown |
| a) Angina | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| b) Myocardial infarct | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| c) Coronary artery bypass grafts/angioplasty | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| d) Recent history of pulmonary edema (i.e. episode(s) of congestive heart failure or pulmonary edema within 6 months prior to dialysis) | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| e) Cerebrovascular disease (i.e. stroke, transient ischemic attack, carotid surgery) | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| f) Peripheral vascular disease (i.e. previous surgery such as femoropopliteal bypass graft, iliac or femoral endarterectomy, angioplasty, etc.; ischemic muscle pain precipitated by exercise; ischemic ulcers; gangrene; amputation) | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| g) Diabetes type 1 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| h) Diabetes type 2 | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| i) Malignancy existing prior to first treatment | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <i>If yes, indicate site using the codes listed on page 3, or specify.</i> | | | |
| _____ | | | |
| j) Chronic obstructive lung disease (i.e. emphysema or chronic bronchitis) | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| k) Receiving medication for hypertension | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| l) Other serious illness that could shorten life expectancy to less than 5 years | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| <i>If yes, specify condition:</i> _____ | | | |
| _____ | | | |
| m) Current smoker (i.e. has smoked cigarettes, cigars or a pipe in the last three months) | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |

CORR study objectives

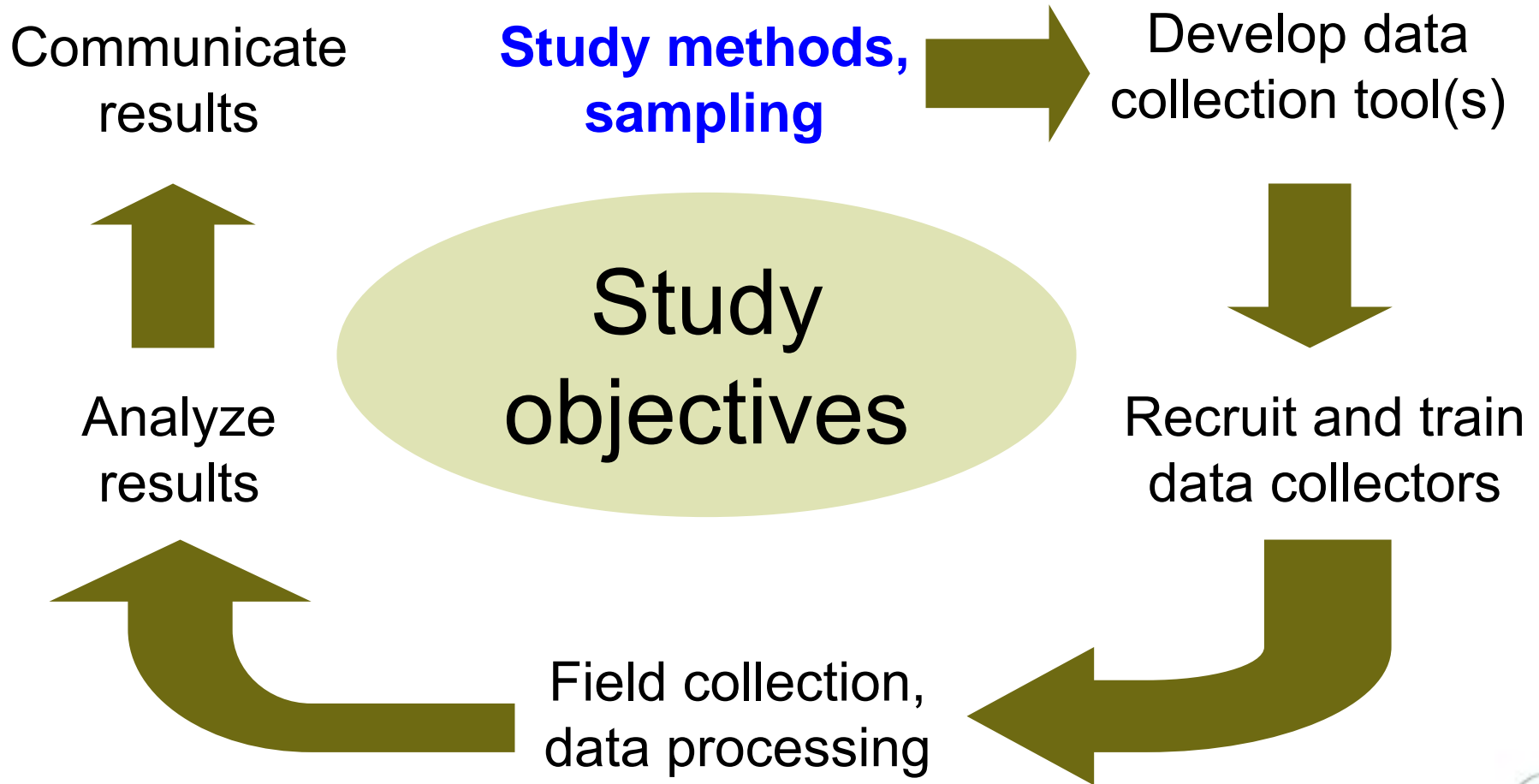
1. **Evaluate the quality of coding of the risk factors** of incident dialysis patients in CORR and also identify the coding issues that arise as a result of the observed coding variation.

2. **Collect information on documentation and data quality processes** from dialysis clinics that register dialysis patients to CORR.

Recoding study

Questionnaire

CIHI data quality studies



CORR recoding study – methods

Objective 1 – evaluate the quality of coding of risk factors

Obtaining “true” values

- The principle concept is to return to **where the data originated.**

Options

- Ask the patient from the dialysis clinic what risk factors they had at the start of treatment. Patient responses are “true” values.
- Perform a chart review at the dialysis clinic, adjudicate discrepancies with clinical staff at the hospital. Adjudicated data are the “true” values.
- Perform a chart review at the dialysis clinic. Data obtained from the chart review are the “true” values.

CORR recoding study – methods

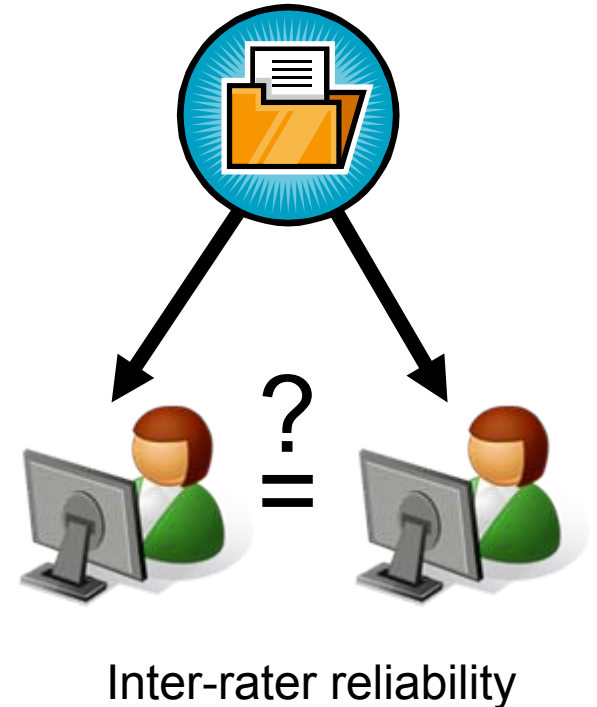
Objective 1 – evaluate the quality of coding of risk factors

Limitation of study method ...

- How reliable are the chart reviewer's data? Are their codes representing "true" values?

How this was addressed ...

- Recruitment/training – to be discussed later
- Data collection – the study included an inter-rater reliability component



CORR recoding study – sampling

Objective 1 – evaluate the quality of coding of risk factors

The sampling methodology was determined by:

1. Asking clients:

- How will the estimates be used?
- How much sampling variance is acceptable in the survey estimates?
- Are estimates required by subgroups of the survey population?

2. Working with a methodologist:

- How much precision is to be gained by increasing the sample size?
- Is it feasible to collect data using the proposed sample given the available budget?



CORR recoding study – sampling

Objective 1 – evaluate the quality of coding of risk factors

- Regional estimates required
- Sensitivity/specificity estimates for each risk factor with defined precision requirements
- Recoding study: two-stage sampling
 - 1st stage, clinics randomly selected
 - 2nd stage, patients sampled within the selected clinics



| | BC | Prairies | Ontario | Quebec | Atlantic |
|---------------------------|----|----------|---------|--------|----------|
| Target pop'n: All clinics | 11 | 13 | 33 | 31 | 11 |
| Clinics sampled | 7 | 7 | 12 | 8 | 5 |

CORR recoding study – sample size

Objective 1 – evaluate the quality of coding of risk factors

Determining sample size needed for the required precision for individual domains

Cochran, Theorem 3.2:

$$V(\hat{p}) = \frac{PQ}{n} \left(\frac{N-n}{N-1} \right)$$

Expressed in terms of the square of the coefficient of variation:

$$CV^2 = \frac{V(\hat{p})}{P^2} = \frac{PQ}{P^2 n} \left(\frac{N-n}{N-1} \right) = \frac{Q}{P} \frac{1}{n} \left(\frac{N-n}{N-1} \right)$$

Solving for n we get:

$$n = \frac{N}{CV^2 \frac{P}{Q} (N-1) + 1}$$

Minimum n for specific domain.

n : sample size

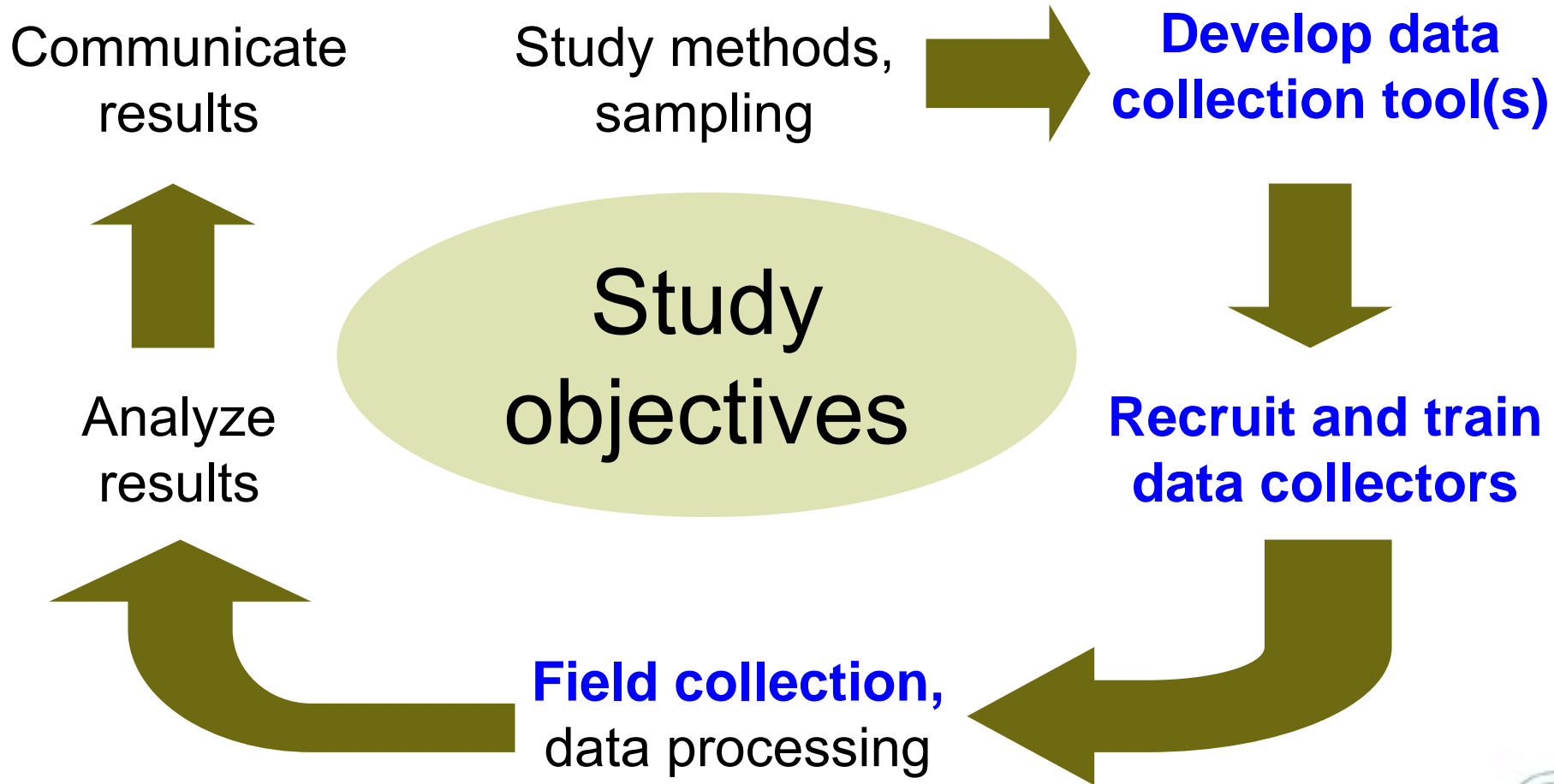
N : population size

P : discrepancy rate (estimated)

$Q = 1 - P$

CV : coefficient of variation

CIHI data quality studies



CORR recoding study – collection tool

Objective 1 – evaluate the quality of coding of risk factors

Step 1

- The chart surveyor reviews the patient chart and enters information into an application.

| Risk factor | | Recoded data | Specific condition identified by chart reviewer | | |
|-----------------------|--|----------------|---|--|--|
| Angina | | Yes | Unstable angina | | |
| Myocardial infarction | | Unknown | N/A | | |
| CABG / angioplasty | | No | N/A | | |

CORR recoding study – collection tool

Objective 1 – evaluate the quality of coding of risk factors

Step 2

- The application then shows the original CORR data and compares original codes to chart reviewer's codes.

| Risk factor | CORR data | Recoded data | Specific condition identified by chart reviewer | Comparison of CORR and recoded data | |
|-----------------------|------------|--------------|---|-------------------------------------|--|
| Angina | No | Yes | Unstable angina | Different | |
| Myocardial infarction | Yes | Unknown | N/A | Different | |
| CABG / angioplasty | No | No | N/A | Same | |

CORR recoding study – collection tool

Objective 1 – evaluate the quality of coding of risk factors

Step 3

- Where different codes are identified, the surveyor records a reason for the difference.

| Risk factor | CORR data | Recoded data | Specific condition identified by chart reviewer | Comparison of CORR and recoded data | Reason for discrepancy |
|-----------------------|-----------|--------------|---|-------------------------------------|-----------------------------|
| Angina | No | Yes | Unstable angina | Different | Chart interpretation |
| Myocardial infarction | Yes | Unknown | N/A | Different | Incomplete doc. |
| CABG / angioplasty | No | No | N/A | Same | N/A |

CORR recoding study – data collectors

Objective 1 – evaluate the quality of coding of risk factors

- **Recruitment** – national campaign with several levels of testing. Successful candidates invited to a training session at CIHI.
- **Training** – focused on (1) coding guidelines for the studied data elements, (2) reviewing patient charts and (3) the study application.
- **Criterion-referenced test** – candidates mimic field collection using real patient charts.
 - Their codes were compared to CIHI's reference standard codes. All deviations were discussed.

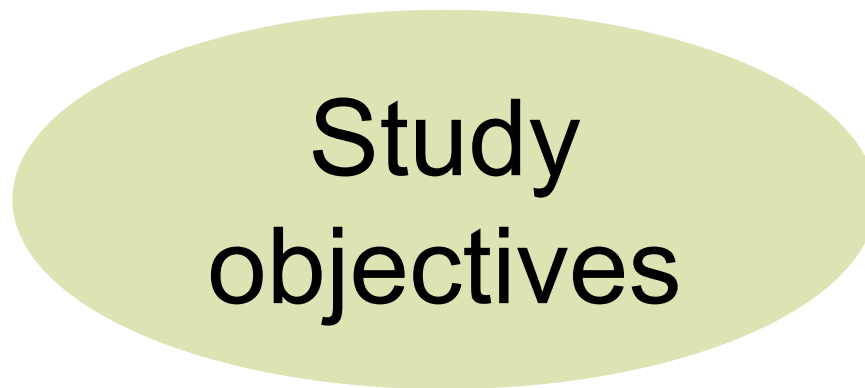


CIHI data quality studies

**Communicate
results**

Study methods,
sampling

Develop data
collection tool(s)

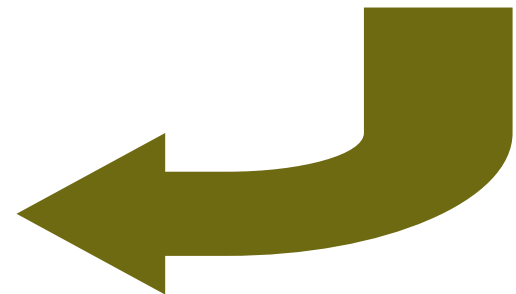


Recruit and train
data collectors

**Analyze
results**



Field collection,
data processing



CORR recoding study – data processing

Objective 1 – evaluate the quality of coding of risk factors

- Editing, data cleaning... and *weighting*.
- Weighting allows findings from the sample of patients to be generalized to the entire patient population.
- CORR study – design weights were adjusted to account for non-response and were calibrated so estimates would sum to known totals.

Design weight for two-stage sample

$$w_d = \frac{1}{\pi_1} \times \frac{1}{\pi_2}$$

w_d : *design weight*

π_1 : *prob(selection at the 1st stage)*

π_2 : *prob(selection at the 2nd stage)*

CORR recoding study – data analysis

Objective 1 – evaluate the quality of coding of risk factors

Estimation

- Weighted estimates only
- Horvitz-Thompson estimators for totals and proportions, bootstrap method for variance

$$\hat{T}_c = \sum_{i,j,k} w_{ijk} \times C_{ijk}$$

Data Analysis for Risk Factors

- Prevalence, overall agreement, Kappa statistic
- Sensitivity, specificity, odds ratios
- Positive and negative predictive values
- Discrepancy reasons

$$\hat{P}_c = \frac{\sum_{i,j,k} w_{ijk} \times C_{ijk}}{\sum_{i,j,k} w_{ijk}}$$

Tests for statistical significance

CORR recoding study – data analysis

Objective 1 – evaluate the quality of coding of risk factors

Gold standard

| CORR database | Data obtained from recoding study | | Total |
|---------------------|-----------------------------------|--------------------|---------|
| | Risk factor present | Risk factor absent | |
| Risk factor present | A | B | A+B |
| Risk factor absent | C | D | C+D |
| Total | A+C | B+D | A+B+C+D |

$$\text{Sensitivity} = \frac{A}{A + C}$$

$$\text{Specificity} = \frac{D}{B + D}$$

CORR recoding study – data analysis

Objective 1 – evaluate the quality of coding of risk factors

Gold standard

| CORR database | Data obtained from recoding study | | Total |
|---------------------|-----------------------------------|--------------------|---------|
| | Risk factor present | Risk factor absent | |
| Risk factor present | A | B | A+B |
| Risk factor absent | C | D | C+D |
| Total | A+C | B+D | A+B+C+D |

$$\text{Positive predictive value} = \frac{A}{A+B}$$

$$\text{Negative predictive value} = \frac{D}{C+D}$$

CORR recoding study – data analysis

Objective 1 – evaluate the quality of coding of risk factors

- Inter-rater reliability
 - How did agreement between the study coders compare to their agreement with the CORR data?
 - Did additional training result in higher agreement between the study coders?
- Comparisons between facility groups
 - Used information obtained via the **questionnaire** to cluster facilities.
 - Performed queries on recoding study data.
 - Do facility practices appear to influence data quality?

CORR recoding study – communication

Objective 1 – evaluate the quality of coding of risk factors

- Staff on the CORR team were
 - involved in planning stages
 - kept informed of progress throughout data collection and processing
 - consulted with during data analysis



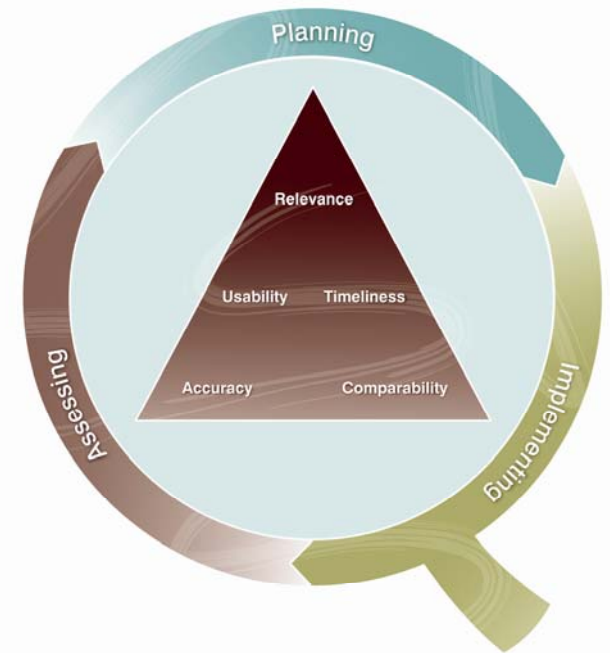
The Goal:

Make this a joint project to increase the uptake of study recommendations.

CORR recoding study – communication

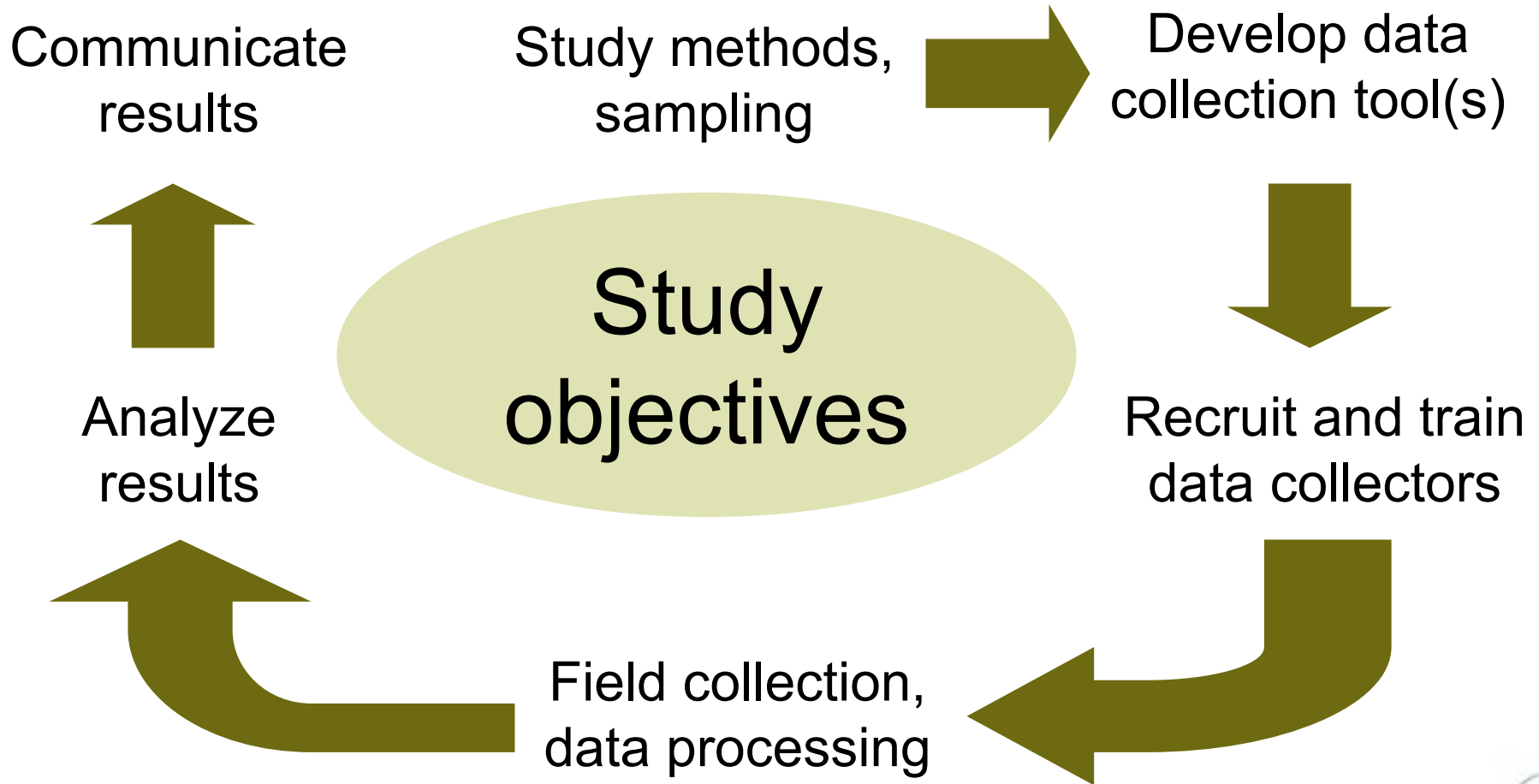
Objective 1 – evaluate the quality of coding of risk factors

- Focus discussion on continuous quality improvement
 - **Assess** data quality
 - **Plan** for improvements
 - **Implement** Q.I. changes



"We can't solve problems by using the same kind of thinking we used when we created them." Albert Einstein

Data quality studies – a summary



In summary...

- Data quality studies are one of many options for assessing data quality and determining the interventions that would improve quality...
- When interventions are implemented, the data should be reassessed to see if the interventions worked.

“One accurate measurement is worth a thousand expert opinions” Grace Hopper

Data Mining at CIHI

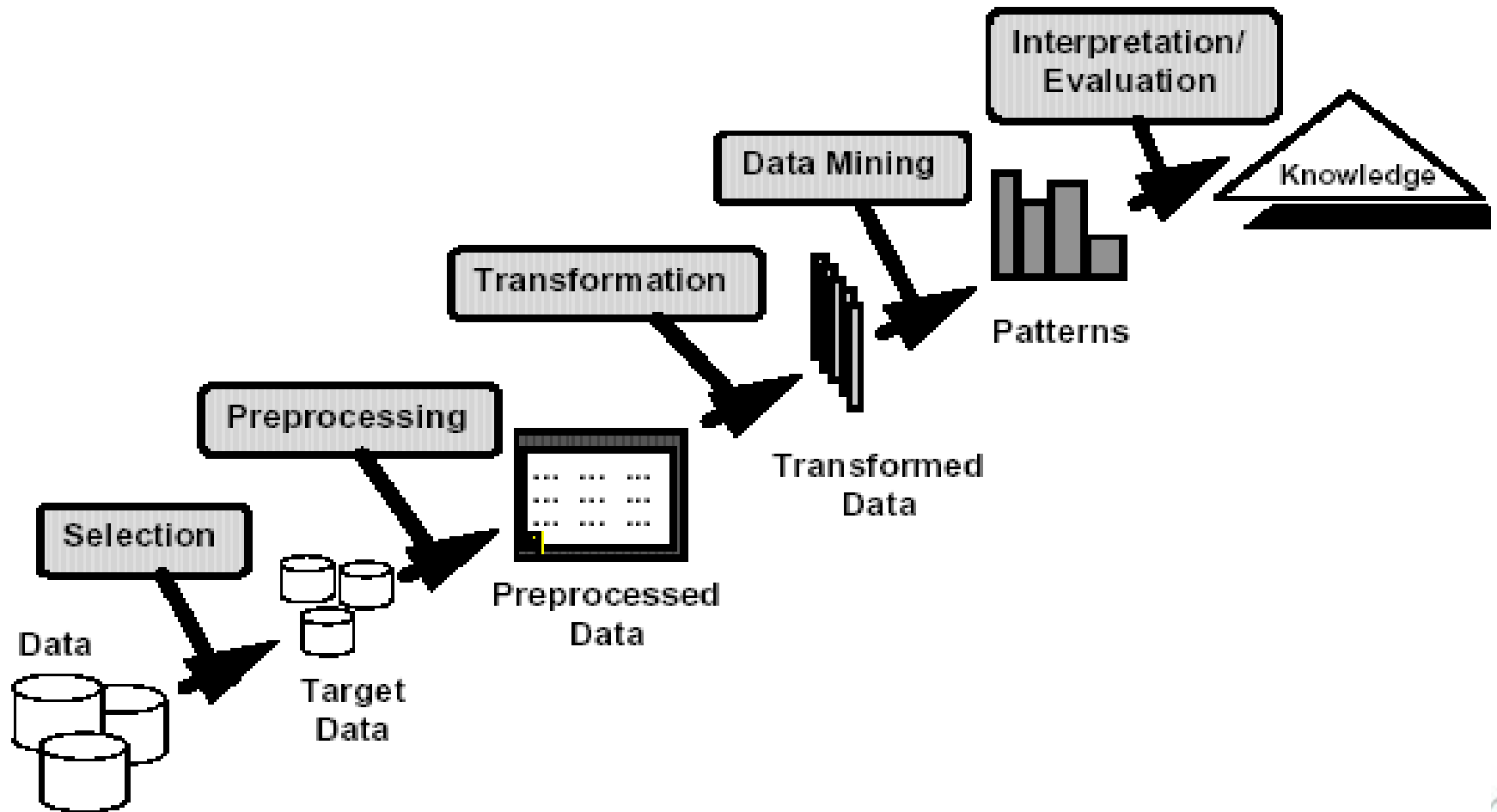
Research and Data Mining Mission Statement

“ To enhance data and information quality at CIHI through:

- Exploratory data analysis of CIHI data holdings
- Statistical consulting
- Statistical education programs
- Promotion of the Data Quality Framework and the services offered by the DQ department

”

What is Data Mining?



Data Mining Methods

- Predictive Methods (supervised learning)
- Descriptive Methods (unsupervised learning)

Research and Data Mining Framework

1. Orientation of DQ research staff to the program area and data holding
2. Initial descriptive scan of data holding
3. Targeted Analyses
4. Exploratory data analysis (EDA)
5. Presentation of Findings
6. Capacity Building

An Example of a Data Mining Project at CIHI

Day Surgery Reporting in Ontario and Nova Scotia

Day Surgery in Ontario and Nova Scotia

- Project started back in April 2008 and completed early July 2008
- The goal of the project was to discern and quantify the impact on volumes of the transition of reporting day surgery from one database to another.
 - Focus on Ontario and Nova Scotia

Day Surgery in Ontario and Nova Scotia

- The transition from one database (DAD) to another (NACRS) included a change in definition of day surgery.
- Severe Acute Respiratory Syndrome (SARS) in Ontario. There was a four-month window where SARS had a significant impact on hospital volumes (both inpatient and ambulatory care visits).
- All happened in the same time frame

Databases Involved

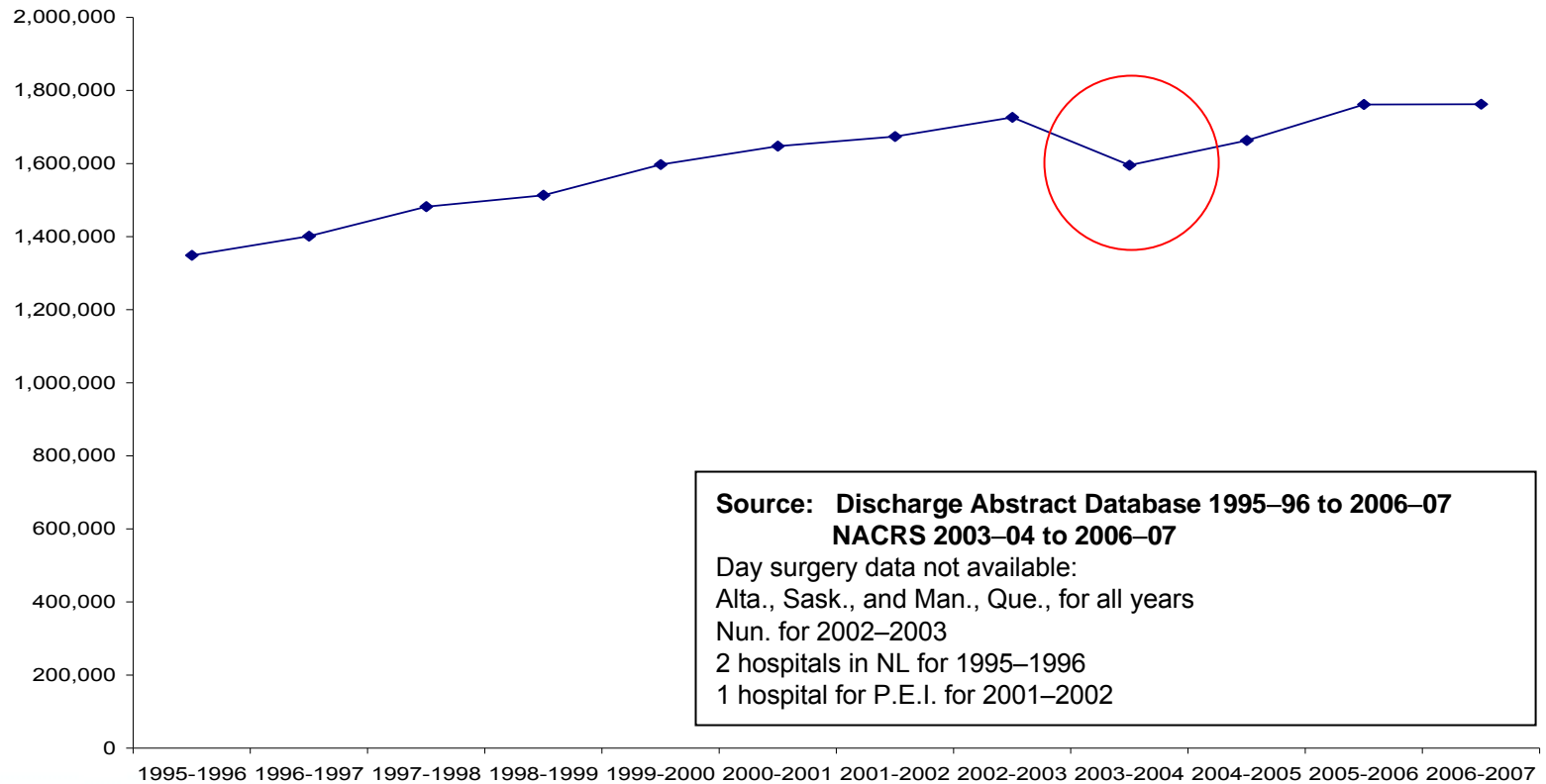
- **The Discharge Abstract Database (DAD)**
 - A national database that contains demographic, administrative and clinical data for acute care inpatient hospital discharges.

- **The National Ambulatory Care Classification System (NACRS)**
 - a database that captures data on client visits to facility and community-based ambulatory care settings.

- **The MIS Standards**
 - provide a standardized framework for collecting and reporting financial and statistical data on the day-to-day operations of health service organizations.

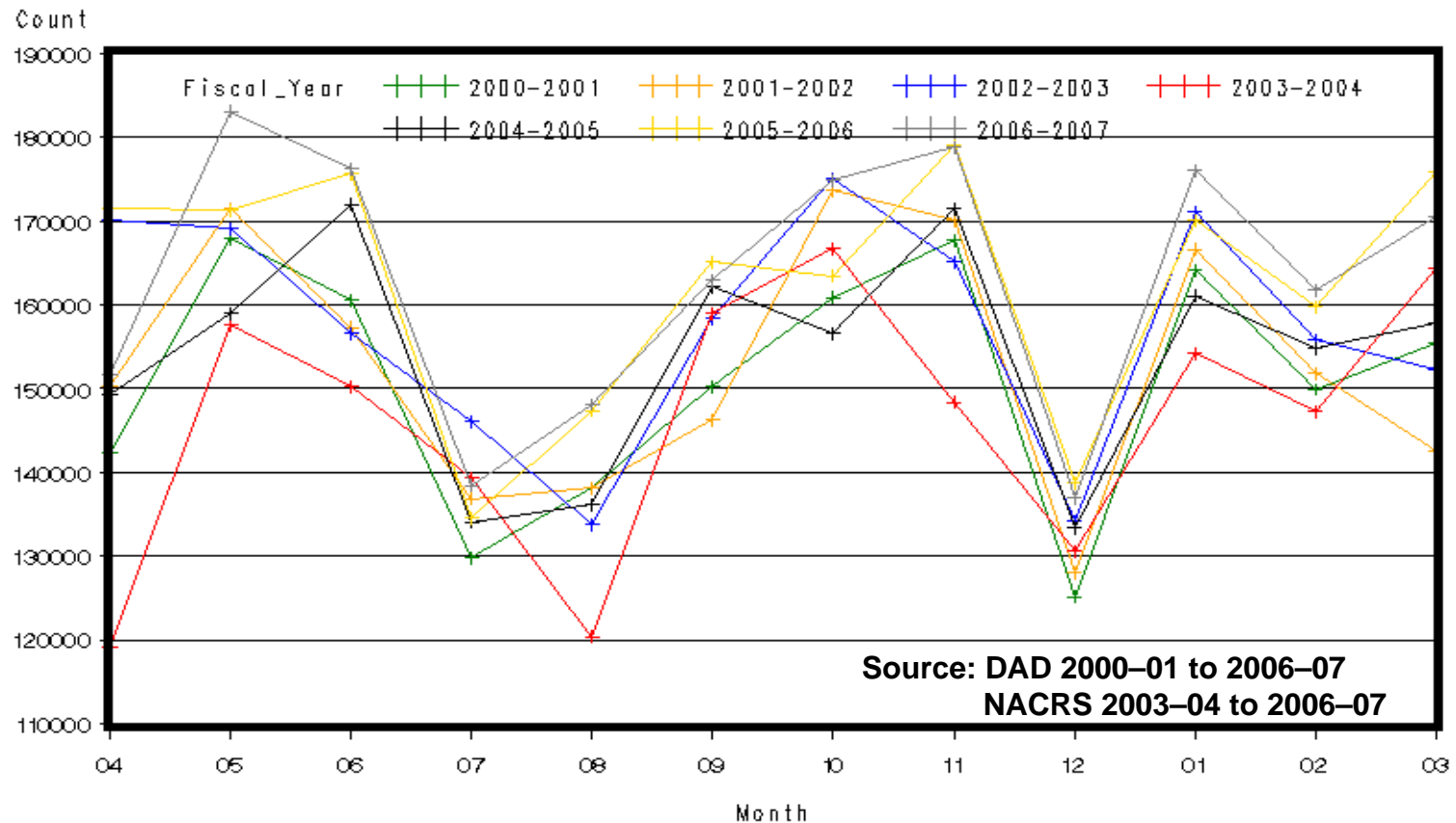
The Impact of the Transition to NACRS of Day Surgery Reporting

Overall trends in day Surgery Volumes in Canada



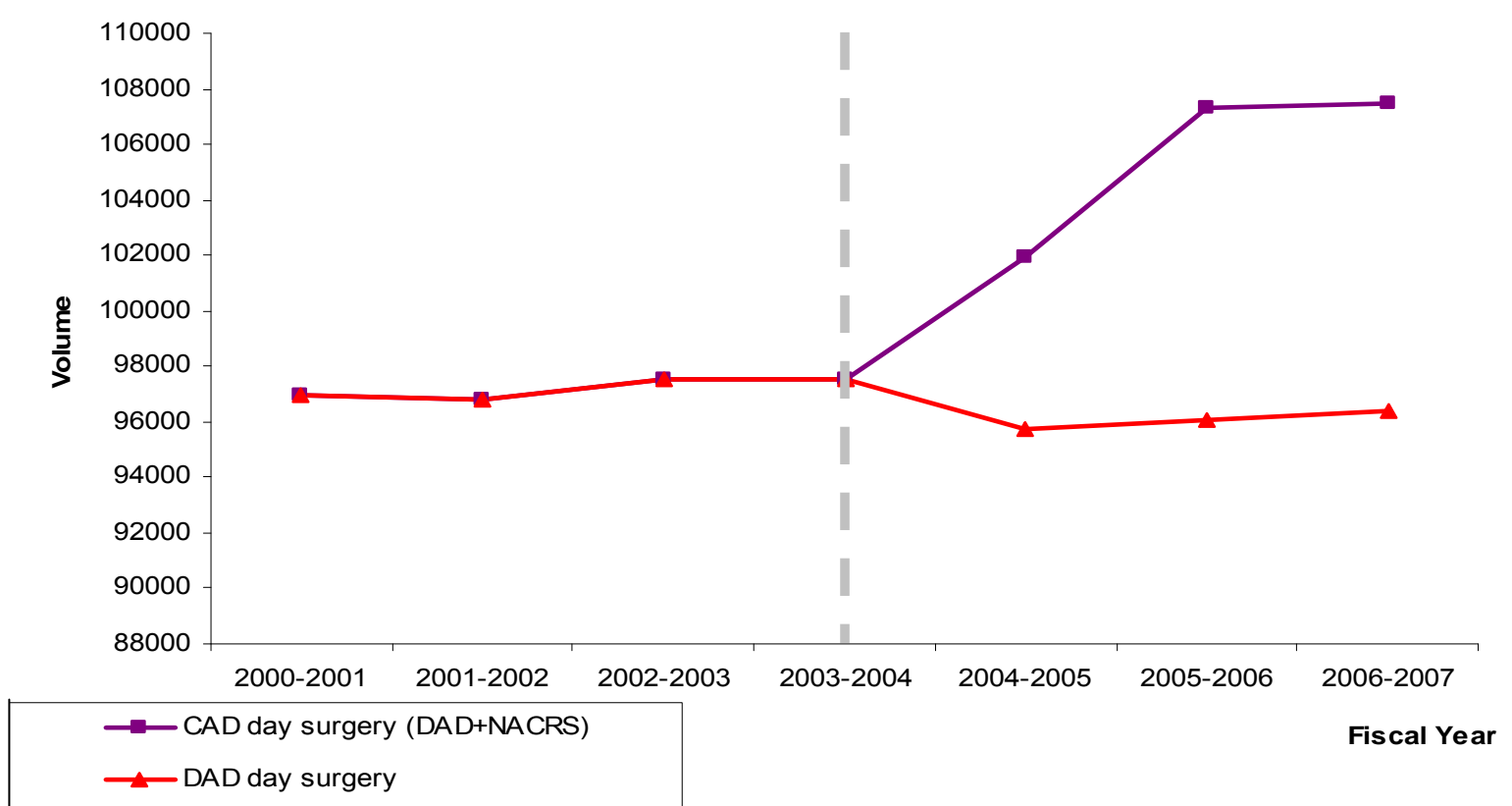
The impact of the transition to NACRS reporting of Day Surgery

Monthly Day Surgery Volumes

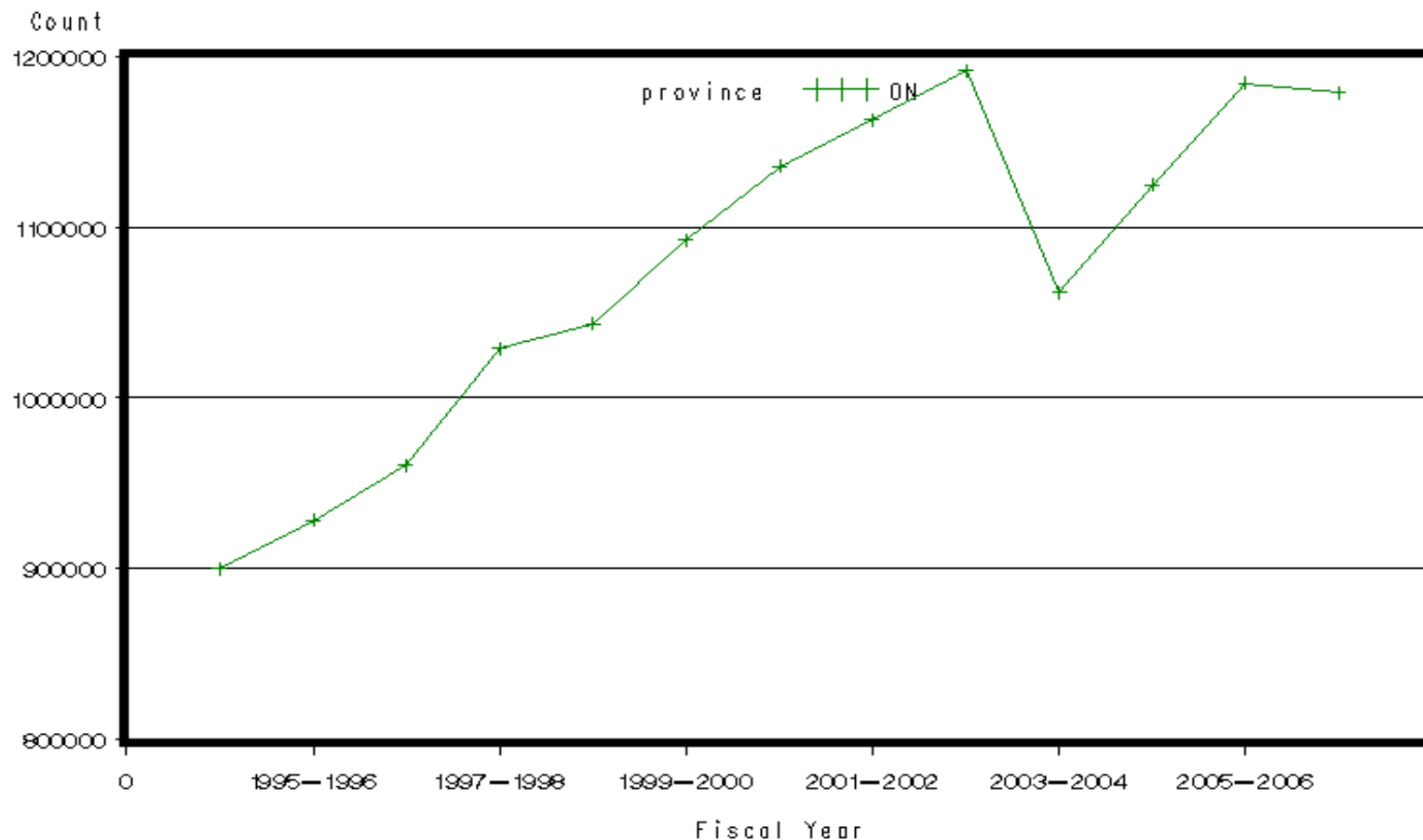


Overall Trends in Day Surgery Volumes by Province -- Nova Scotia

Day Surgery Volumes in Nova Scotia by database



Overall Trends in Day Surgery Volume by Province -- Ontario



Some background about Auto-Regressive of order P, AR(P) model

- The general formula for Auto-regressive model is as follows:

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t$$

Where

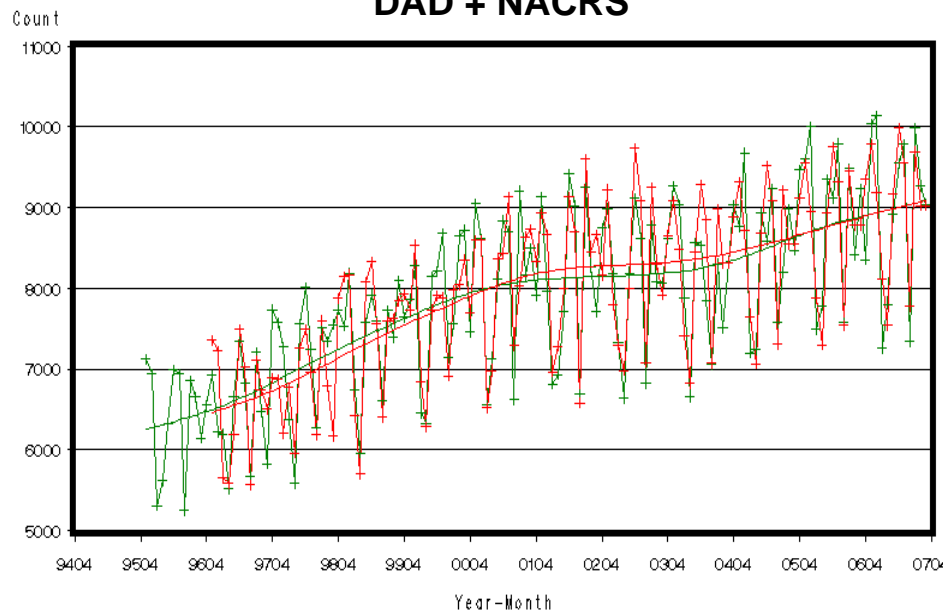
- $\varphi_1, \dots, \varphi_p$ are the parameters of the model and c is a constant and ε_t is white noise
- The absolute value of φ_i , must be less than 1
- $\hat{\mu} = \frac{c}{1 - \hat{\varphi}}$
- Similarly, $Var(X_t) = \frac{(c + \varphi\mu)^2 + \sigma^2 - \mu^2}{(1 - \varphi)^2}$

Time Series Model on Day Surgery Impact Analysis

- A simple autoregressive model used to forecast monthly DS Volumes.
- 1995-2002 data used for modeling and forecasting 2003-2006 volumes
- An AR(3) model fits and predicts Nova Scotia data well, but not Ontario:

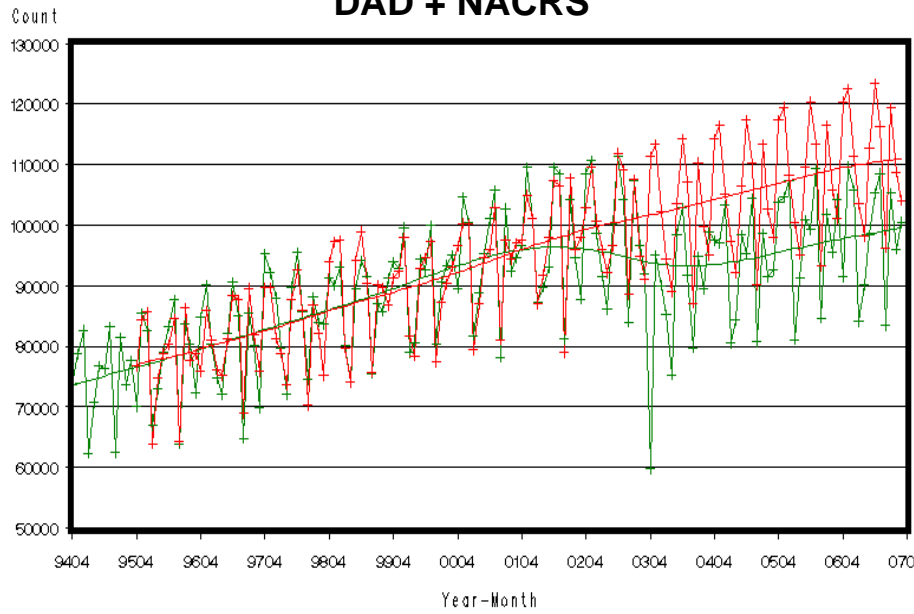
Forecast vs. Real Count in N.S.

DAD + NACRS



Forecast vs. Real Count in Ont.

DAD + NACRS



Intervention Models

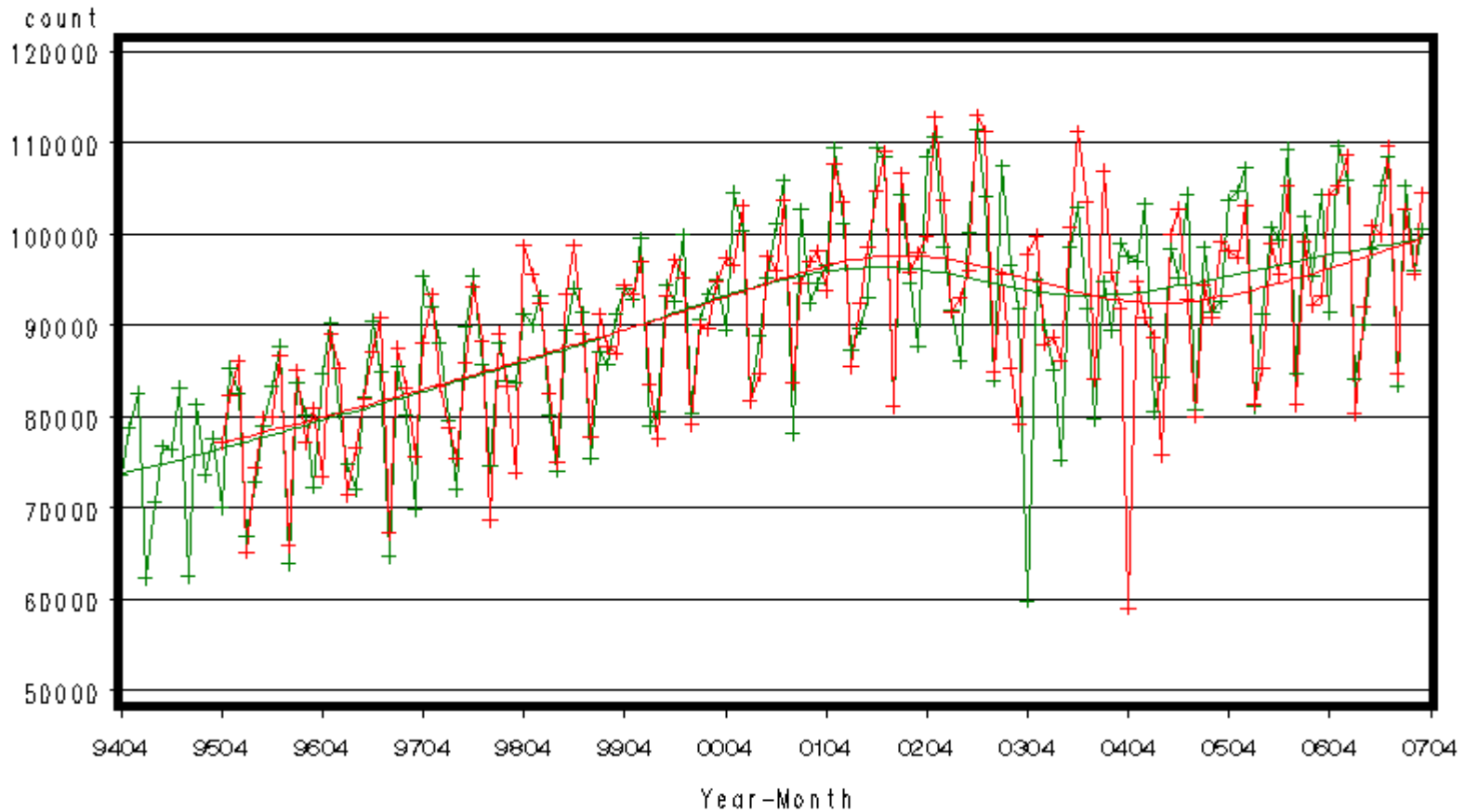
- A special autoregressive model with input series, also called *interrupted time series* model
- Used for forecast and analyze the impact of the intervention
- *Impulse vs Continuing Interventions*
- The model has the form of:

$$X_t = \mu + \sum_{i=1}^p a_i X_{t-i} + b_{SARS} I_{SARS} + b_{Trans} I_{Trans} + \varepsilon_t$$

Where, X_t is the monthly DS volume, a_i, b_j are the model parameters, I_{SARS} , I_{Trans} are the indicator variables and ε_t is the model error. μ is constant value.

Results of Intervention Model

Forecast vs Real Count in Ont., Intervention Model



Conclusions from Day Surgery Project

- Overall, the goal of this project was to discern the impact of the transition of day-surgery reporting from DAD to NACRS.
- Once SARS was accounted for, there was no statistically significant impact on day-surgery volumes during the transition period.
- The change in definition of day surgery did not significantly affect overall volumes of day surgeries reporting from DAD to NACRS.

Data and Information Quality – CIHI a Shared Responsibility ICIS

Together, we engage in rigorous activities to ensure data collected and provided are of the highest standard





Data and Information Quality

The foundation for meaningful health information

Thank you!

www.cihi.ca / www.icis.ca

References

Journals and Conference Proceedings on DQ Special Studies

- Gibson D. et al, (2008) 'The National Ambulatory Care Reporting System: Factors that Affect the Quality of its Emergency Data', Int. J. Information Quality.
- Landis J. and Koch G., (1977) 'The Measurement of Observer Agreement for Categorical Data', Biometrics.
- Longenecker et al, (2000) 'Validation of Comorbid Conditions on the End-Stage Renal Disease Medical Evidence Report: The CHOICE Study', Journal of the American Society of Nephrology.
- Merkin S. et al (2007) 'Agreement of self-reported comorbid conditions with medical and physician reports varied by disease among end-stage renal disease patients', Journal of Clinical Epidemiology.
- Quan H. et al, (2002) ' Validity of Information on Comorbidity Derived From ICD-9-CCM Administrative Data', Medical Care.
- Richards H. et al, (2007) 'Assessing Data Quality Using a Complex Study Design', Proceedings of the Australasian Conference on Information Quality.
- Richards H. et al, (2006) 'Re-abstraction Study of Discharges from Ontario Case Costing Hospitals', Proceedings of the International Conference on Information Quality.
- Williams S. et al, (2006) 'Assessing the Reliability of Standardized Performance Indicators', International Journal for Quality in Health Care.

Textbooks on Survey Sampling and Data Mining

- Cochran, E.G, Sampling Techniques (United States of America: John Wiley & Sons Inc., 1977).
- Devore, J.L., Probability and Statistics for Engineering and the Sciences (Duxbury Press. 1995).
- Fayad, Usama M. et. al, Advances in Knowledge Discovery and Data Mining (The MIT Press. 1996)
- Lohr, S.L., Sampling: Design and Analysis (Duxbury Press. 1999).
- Scheaffer, R. et al, Elementary Survey Sampling (Duxbury Press. 1996).
- Statistics Canada, Survey Methods and Practices (Minister of Industry. 2003).